Guiding Supervised Learning by Bio-Ontologies in Medical Data Analysis

Janusz Wojtusiak, Hua Min, Eman Elashkar, Hedyeh Mobahi

Guiding Supervised Learning by Bio-Ontologies in Medical Data Analysis

Janusz Wojtusiak, Hua Min, Eman Elashkar, Hedyeh Mobahi

Health Informatics Program, George Mason University, Fairfax, Virginia, USA

{jwojtusi,hmin3,eelashka,hmobahi2}@gmu.edu

Abstract. Ontologies are popular way of representing knowledge and semantics of data in medical and health fields. Surprisingly, few machine learning methods allow for encoding semantics of data and even fewer allow for using ontologies to guide learning process. This paper discusses the use of data semantics and ontologies in health and medical applications of supervised learning, and particularly describes how the Unified Medical Language System (UMLS) is used within AQ21 rule learning software. Presented concepts are illustrated using two applications based on distinctly different types of data and methodological issues.

Keywords: Supervised Machine Learning, Biomedical Ontologies, UMLS

1 Introduction

Recent advancements of Machine Learning (ML) made it applicable to wide range of problems, including those in medical, healthcare and health domains. These methods are able to make accurate predictions in uncertain environments, by finding patterns scattered over massive amounts of data. Strength of many of the newest methods comes also in the ability to combine Natural Language Processing (NLP) tools with learning from structured data. The majority of novel methods are statistical and focus on analysis of numeric data. The principles of these machine leaning methods usually rely on distributions and patterns inside the data sets only. They do not include the meanings of the data sets. Domain knowledge for such methods is limited to ad-hoc encoding of attributes in the data or prior parameters of the model being learned as in examples of deep learning of neural networks [1]. Surprisingly, very few machine learning methods allow for modeling of domain knowledge in order to guide the learning process which can potentially make the learned models closer to human decision-makers.

The presented research explores the utilization of ontologies and data semantics to guide machine learning process. It provides the additional information to those distributions and patterns already inside the data sets. The motivations of this research come from the existence of rich biomedical ontologies created for the medical data integration methods, NLP algorithms and the Semantic Web [2], as well as early research on human-oriented machine learning. This presence of domain knowledge provides an ideal

opportunity to complement pure data for machine learning, with relationships that span over the data.

The researchers started to utilize domain knowledge to guide machine learning in some fields. For example, the frequent utilizations of the ontology exist in NLP, including ontology-based methods for indexing, extracting, and analyzing clinical notes [3,4,5]. In [6], the researchers classified patients with different types of epilepsy using different methodologies including ontology-based classification (OBC). The OBC achieved better results than others did.

After the brief introductions for biomedical ontologies and supervised machine learning, we will present an approach to handling semantic information and reasoning with ontologies in supervised learning.

1.1 Biomedical Ontologies

An ontology formally represents domain knowledge as a set of concepts and relationships between those concepts. The concepts in an ontology should be close to objects (physical or logical) in the real world. Relationships describe the interactions between concepts or a concept's properties. The most important relationship is the "IS-A" hierarchical relationship. It serves as the ontology's backbone and supports the property inheritance. The "IS-A" relationship connects a more specific concept (child concept) to a more general concept (a parent). Non-IS-A relationships, called associative or semantic relationships, connect concepts across the hierarchies in an ontology. Broadly speaking, ontologies include thesauri, terminologies, classifications, and coding systems. Ontologies play important roles in biomedical research through a variety of applications including data integration, knowledge management, natural language processing, and decision support [7].

The most popular biomedical ontologies (Bio-Ontologies) include Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT) [8], International Classification of Diseases (ICD) [9], Logical Observation Identifiers Names and Codes (LOINC) [10], Gene Ontology (GO) [11], Medical Subject Headings (MeSH) [12], RxNorm [13], Foundational Model of Anatomy (FMA) [14], and National Cancer Institute Thesaurus (NCI Thesaurus) [15]. Each ontology has its own purpose and scope. For example, SNOMED CT is a systematically organized computer processable ontology of medical terms. ICD defines the universe of diseases, disorders, injuries and other related health conditions. LOINC is a coding system for laboratory and clinical observations. GO provides controlled vocabularies of defined terms representing gene product properties including cellular components, molecular functions and biological processes. MeSH is designed to provide a hierarchically-organized terminology for indexing and cataloging of biomedical information such as MEDLINE/PubMed. RxNorm, published by National Library of Medicine (NLM), provides normalized names and a model for clinical drugs available in the US. FMA represents a coherent body of explicit declarative

knowledge about human anatomy. Finally, NCI Thesaurus includes broad coverage related to the cancer research domain.

Therefore, there are communication barriers between various information systems or applications if the developers use different vocabularies in different systems. In order to solve these barriers, the Unified Medical Language System (UMLS) was developed by the NLM in 1986 [16] and it is constantly being updated. The UMLS has three knowledge sources: Metathesaurus, Semantic Network, and SPECIALIST Lexicon. The UMLS (2016AB) contains more than 3.44 million concepts (Concept Unique Identifiers; CUIs), 22 million relationships among those concepts, and 13.7 million unique concept names (AUIs) from 199 source vocabularies. One important goal of the UMLS is to establish mappings between different Bio-Ontologies. A concept unique identifier (CUI) is assigned to the terms from various source ontologies that have the same meaning in the Metathesaurus. The mappings among these vocabularies allow computer systems to translate data among the various information systems. The rich relationships in the UMLS also provide a solid foundation for reasoning in the medical knowledge [7]. Other UMLS applications include providing browser for its source ontologies, the Clinical Observations Recordings and Encoding (CORE) Problem List Subset [17], NLP [3,4,5], and value sets for Clinical Quality Measures (CQMs) [18].

1.2 Supervised learning

While machine learning is a broad area, the presented work is focused on supervised learning, and more specifically concept learning. The methodology described here can be applied to output concepts which are independent, ordinal or structured. One can also extend the method presented here into regression learning, as well as other forms of machine learning, i.e. unsupervised and reinforcement learning.

The problem considered here is to learn a model $M: X \rightarrow Y$ which can be viewed as a function that assigns classes from $Y=\{y_j\}$, $j=1..k$ into objects from X. Learning is performed by an algorithm A given dataset D and background knowledge BG, $A(D, BG) = M$, where $D=\{(x_i, y_i), i=1..N\}$. This work focuses on the background knowledge, BG, and specifically its forms that can be retrieved from ontologies.

Many concept learning methods have been developed in the field, including symbolic methods for learning decision trees [19] or rules [20], numeric methods for learning sets of equations such as Support Vector Machines [21], Logistic Regression [22], or non-Negative Matrix Factorization [23]. Regardless of the used method, the general goal is to build models that maximize quality measure (or minimize loss function). An issue considered here is how to improve the methods when additional structure about the problem is known, i.e., in the form of hierarchical relationships between values of attributes, or non-IS-A relationships between attributes. In a sense, recent work on neural networks, related to deep learning [24] can be viewed as a form of encoding of problem into structure of model. In deep learning different structures of networks are

considered, which can be based on hierarchies, but typically represent components of the problem rather than semantic concepts.

Early and more recent work on learning structures [25] and use of background knowledge included advanced in Inductive Logic Programming (ILP) and hierarchical learning methods. In the ILP, the background knowledge is defined as a set of relations (predicates) that can be used in the definition of the target concept [26]. An ILP system derives rules based on an encoding of the known background knowledge and a set of examples represented as a logical database of facts. Another related field is the hierarchical Relational Reinforcement Learning (HRRL) [27,28,29]. In those studies, hierarchies have been used to reduce the complexity of decision making and improve the actual process of learning. Both ILP and HRRL are particularly useful in bioinformatics, healthcare, and NLP.

1.3 Example Data

This paper illustrates concepts of using ontologies and semantics in machine learning on two examples of medical/health data. These are chosen because of inclusion of diverse types of data and are part of existing projects on using semantics in machine learning by the authors.

- **SEER-MHOS**: The example learning problem concerns the ability to automatically assess patient disabilities in performing Activities of Daily Living (ADLs). Such activities are important measures of patient independence, quality of life and need for care. However, the data about ADLs is not routinely collected along with clinical or administrative data. The purpose of this application is to automatically assess patients' functional disabilities based on general demographic information and known broad categories of diagnoses. Models are trained on SEER-MHOS (Surveillance, Epidemiology, and End Results – Medicare Health Outcomes Survey) which is a linked dataset. A subset of data with 1,849,311 unique patients, out which 102,269 patients diagnosed with cancer, have been used for this research. SEER is a cancer registry program that provides clinical, demographic and cause of death information [30]. MHOS data is a survey based report that contains both patient conditions and ADLs. In this research, SEER-MHOS data has been coded with UMLS CUIs and used to create models for predicting ADLs after cancer diagnosis.

- **MIMIC-III:** Learning from clinical data adds another level of complexity beyond standard administrative healthcare data. The MIMIC III ('Medical Information Mart for Intensive Care') is a large database that includes de-identified, comprehensive clinical data of patients admitted to critical care units at a large tertiary care hospital, Beth Israel Deaconess Medical Center in Boston, Massachusetts. The data are publicly available to researchers who satisfy certain conditions [31]. MIMIC-III consists of over 58,000 hospital admissions for 38,645 adults and 7,875 babies. It is structured into 26 tables organized as a relational database. Data have been collected during routine hospital care between 2001 and 2012 and was downloaded from several

sources including archives from critical care information systems, hospital electronic health record and Social Security Administration Death Master File [32]. In the presented work, MIMIC-III data has been mapped to UMLS ontology and used to create models for predicting 30-day post-hospitalization mortality.

2 AQ21, Semantics, and Ontologies

AQ21 is the latest of rule learning systems developed by Ryszard Michalski's team at George Mason University and previously at University of Illinois [33]. Currently AQ21 is being extended by methods that allow for reasoning with complex data, i.e., data that is mapped into ontologies, data with multiple types, and specifics of medical and health data [34]. The AQ family of rule learners follow traditional separate-and-conquer approach to learning, by generating multiple stars (all rules one positive example, *the seed*, that do not cover negative examples). From each star top rules are selected for further processing. This operation is repeated until all positive examples are covered in the training data. Finally, the learned rules are optimized to maximize their quality according to user-defined criteria. AQ21 software implements several additional modules for adjusting representation space by constructive induction [35,36], testing and applying data, handling time series, generating natural language descriptions from rules, and others. The research on AQ rule learners follows the idea of *natural induction* [37] in which created models are in forms natural to people (transparent and consistent with their prior knowledge).

The general principle behind the work presented here is that a rule learning system that understands semantics as well as relationships between attributes or values, can reason better than one that is provided only data. This principle is grounded on how people reason. Instead of purely relying on data, people use their knowledge to put all data in context and reason about it. For example, knowing that Type I Diabetes and Type II Diabetes are both Diabetes, reduces complexity and allows for the learned description to be more general (handling IS-A relationships and reasoning with hierarchies is described later in this chapter). By semantics of data, we understand attribute types, meta-values, aggregated vs. individual data, and relationships between data elements.

Attribute types are used to specify which methods of reasoning, including data transformations, can be applied when learning. The basic recognized attribute types are: nominal (unordered sets of symbolic values, i.e., *[treatment=radiation, surgery]*), ordinal (ordered sets of symbolic values, i.e., *[stage=I..III])*, cyclic (ordered sets of symbolic values that form a cycle, i.e., *[day=Friday..Monday]*), structured (hierarchical sets of symbolic values, i.e., *[treatment=surgery]* with surgery having subcategories such as robotic surgery, etc.), graph (values are linked together by edges in a graph, i.e.,), set (multiple values can be selected at the same time, i.e., *[diagnosis={diabetes, hypertension, obesity}]*), set-defined (similar to set, but with additional structure added

on top of values, i.e., diagnoses in previous example form a hierarchy), interval (numeric values with defined addition and subtraction, zero is not defined), ratio (numeric values with defined multiplication and division), cyclic-ratio (numeric values forming a cycle, i.e., *[angle=276..15]*), and absolute (numeric values with only order defines and no operations permitted, i.e., social security number). These attribute types along with additional examples are explained in [38]. Attribute types are critical when attempting to generalize and reason with rules, as well as apply constructive induction methods, i.e. interval attributes should not be multiplied.

Aggregated data refer to data in which examples provided to the learning system describe a group of individual observations, rather than an individual object about which the system reasons [39]. Aggregated data are typically described using mean and standard deviation of attributes measured for a group of examples, or frequency of symbolic values. The learning problem from aggregated data is to create models for categorizing individual data when no individual training data are available, or only small portion of individual data are available with addition of aggregated data. Learning from aggregated data is particularly important in fields such as healthcare in which access to individual patient data is difficult or impossible. It is inspired by the field of meta-analysis in which aggregated data retrieved from published scientific studies can be analyzed to arrive at global conclusions supported by majority of studies.

Meta-values refer to special values present in the data, namely unknown, not applicable and irrelevant [40]. These meta-values correspond to potential reasons for which regular values are not present: they are not known, do not make sense, or are removed based on expert's judgment. The majority of machine learning and data mining methods ignore the fact that not all missing data are the same. Imputation methods can be meaningfully applied only to values that exist but are not known. Imputing data for not-applicable values simply does not make sense (for example replacing missing Prostate-specific Antigen (PSA) missing value for female patients in medical dataset). Similarly, imputing irrelevant values that were deliberately removed by experts makes no sense. In statistics, there is a distinction between data missing at random, missing completely at random, and missing not at random. While these classes partially correspond to semantic meaning of meta-values their interpretation is different. Meta-values represent background knowledge that is handled internally within learning systems such as AQ21.

Medical data in the electronic health record systems (EHRs) include a wide range of data from patient demographic, medical history, diagnosis, treatments, socioeconomic status, to genetic information. The medical data can be coded with concepts from one or multiple medical ontologies. Those concepts are connected by different type of relationships including most typically used IS-A relationships being part of hierarchies, or non-IS-A relationships that carry semantic meaning of the connections between concepts. The concepts and their relationships are represented in the medical ontologies in formal knowledge representation languages such as OWL and OBO. Thus, they are computationally processable by machine learning methods. In summary, semantic data

description includes information about attribute types, inter-attribute relationships, value aggregation semantics, data transformations, and meta-values.

The main focus of the presented work is to use the semantics of data when applying supervised machine learning methods to construct classification models. The following sections describe how relationships can be extracted from UMLS and included as part of background knowledge used by AQ21. Generalization with hierarchies and IS-A relationships, learning with non-IS-A relationships, and finally learning hierarchical and ordinal models by AQ21 are presented.

2.1 Hierarchy Extraction from UMLS

The data pre-processing includes one major step that is a multi-step hierarchy extraction. It contains 4 steps: (1) Mapping data to the UMLS concepts (and identify their CUIs); (2) Extracting complete sub-hierarchy by following IS-A relationships using CUIs from step 1; (3) Resolving inconsistencies in the hierarchy (e.g., cycles, duplicates); and (4) Encoding extracted hierarchies in ML-software readable format (i.e., AQ21 requires a list of parent-child pairs for all relationships that form hierarchy). The detailed description is outlined below:

1. Map data to the UMLS concepts. The mapping is a challenge since the meaning of the concepts varies depending on specific sources and authors. For example, the definition of Congestive Heart Failure (CHF) from Elixhauser et al. in 1998 [41] is different from UMLS. According to the definition from Elixhauser, there are 18 ICD-9 codes that are classified as CHF (see first column in Table 1). While in UMLS, to understand how CHF is defined, we need to follow hierarchy starting from the most general Congestive Heart Failure (CHF) concept, tracking down all its possible children through the hierarchy. This was done by tracking children of the concept, then children of all children, to last concept related to the main general concept, as shown in Figure 1. However, there are only 11 concepts (CUIs) that defines Congestive Heart Failure (CHF) in the UMLS system (and corresponding ICD-9 codes) as shown in second column in Table 1. In this case, CUIs resulting from tracking hierarchy of CHF concept in UMLS, were mapped again to corresponding ICD9 so we can show concepts in both Elixhauser and UMLS defined using ranges of ICD9. The seven inconsistent codes are highlighted in gray in Table 1. This is just one example of the fact that the mapping process is difficult, and currently needs to be done manually by domain experts.

2. Follow IS-A relationships in the UMLS for both parents and children until complete sub-hierarchy is extracted. The hierarchy should extend to the farthest common ancestors and descendants. In UMLS one should follow relationships corresponding to all considered source terminologies, not only one in which the original data is coded.

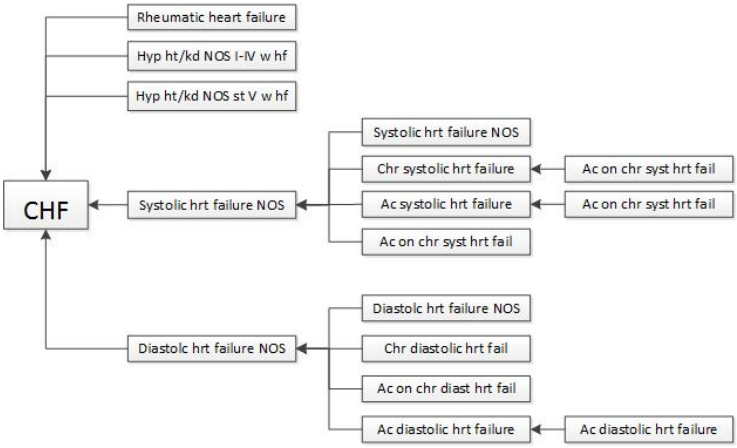


Fig. 1. Congestive Heart Failure concept hierarchy extracted from UMLS.

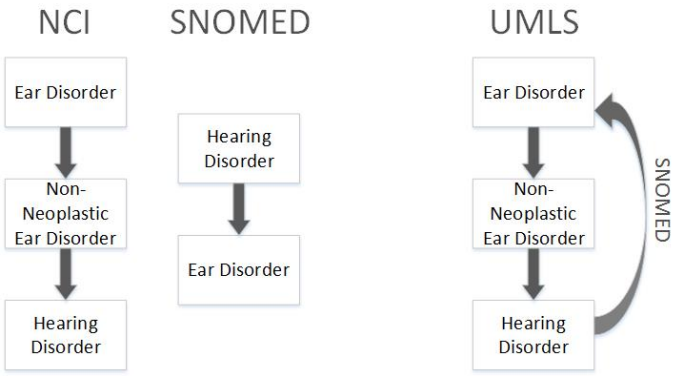


Fig. 2. Example cycle in UMLS created when combining multiple terminologies (SEER-MHOS Data).

3. Resolve inconsistencies in the extracted hierarchies. Due to nature of UMLS and fact that it is constructed from multiple source terminologies, a number of inconsistencies may happen. For example, the extracted hierarchies often include cycles, which are not permitted in AQ21 (and make it impossible to reason with data). Figure 2 shows one example of cycles in UMLS. Cycles are resolved by breaking links that go back to concepts higher in the hierarchy, as measured by distance from the UMLS root. Other types of inconsistencies include use of duplicate concepts or depreciated concepts. Those inconsistencies should be removed from the final hierarchy.

4. Encode hierarchy in ML-software readable format, typically a list of parent-child pairs. AQ21 is a stand-alone software that reads input from text files. The files need to include all semantic information required to correctly reason with the data. Specifically, in AQ21 hierarchical relationships are part of definition of domains attributes that describe data.

Currently the steps 1-3 are completed outside of AQ21 software, but the system is being extended by the ability to link directly to UMLS. In principle, with right data pre-processing and representation transformations, it may be possible to encode semantic information in a way that most ML methods can use it.

Table 1. Congestive Heart Failure (CHF) as defined by Elixhauser et al. and in the UMLS hierarchy (MIMIC III Data).

Elixhauser Definition (ICD-9)	UMLS Definition (CUI)	Description
398.91	C0155582	Congestive rheumatic heart failure
402.01	Not CHF	Malignant hypertensive heart disease W heart failure
402.11	Not CHF	Benign hypertensive heart disease W heart failure
402.91	Not CHF	Unspecified hypertensive heart disease W heart failure
404.01	Not CHF	Hypertensive heart & chronic kidney disease, malignant, w heart failure & chronic kidney disease stage I - stage IV, or unspecified
404.03	Not CHF	Hypertensive heart & chronic kidney disease, malignant, w heart failure & chronic kidney disease stage V - end stage renal disease
404.11	Not CHF	Benign hypertensive heart & renal disease W CHF
404.13	Not CHF	Benign hypertensive heart & renal disease W CHF & renal failure
404.91	C3665458	Hypertensive heart & renal disease W heart failure & renal failure
404.93	C0494576	Heart Failure, Systolic
428.20	C1135191	Acute systolic heart failure
428.21	C2732748	Chronic systolic heart failure
428.22	C1135194	Acute on chronic systolic heart failure
428.23	C2733492	Heart Failure, Diastolic
428.30	C1135196	Acute diastolic heart failure
428.31	C2732951	Chronic diastolic heart failure
428.32	C2711480	Acute on chronic diastolic heart failure
428.33	C2732749	Hypertensive heart & renal disease W heart failure & renal failure

2.2 Generalization with Hierarchies

Existing methods in AQ21 allow for using hierarchies extracted from ontologies to be used in specifying optimal generalization level of rules. The method is based on an approach first described by Kaufman and Michalski [42] and implemented in earlier AQ systems. The method follows IS-A relationships in the data to generalize or specialize rules in order to improve their accuracy and simplicity as defined by implemented rule quality measures.

The method can be applied to generalize beyond a single attribute in the data that implements set-valued attributes using binary indicators. This is particularly useful when analyzing coded medical data with potentially hundreds of thousands of binary attributes. For example, patient diagnoses coded using ICD-9 codes can result in the need to create close to 10,000 binary attributes. It is important to generalize ICD-9 codes in order to reduce the number of features. The generalization can be done by categorizing ICD-9 codes into Clinical Classifications Software (CCS) codes or finding a common ancestor for those codes by climbing the UMLS hierarchy. In this study, CCS was applied to generalize ICD-9 codes for the SEER-MHOS dataset. The goal was to find the predictor or set of predictors for the ADLs deficiencies.

Before analysis, the data were preprocessed as follows: First we limited our study population to those patients who completed at least one survey before their cancer diagnosis and one survey roughly one year after the diagnosis. If a patient had multiple surveys, we used the surveys closest to before and after the cancer diagnosis. These very strict criteria significantly reduced the data size and the process produced a cohort of 723 cancer patients. The set of output attributes included six ADL indicators (walking, dressing, bathing, moving in/out chair, toileting, and eating) that were extracted from the survey completed after the cancer diagnosis. Input attributes, extracted from survey completed prior to cancer diagnosis and cancer registry, were based on known ADL factors from the literature [43,44,45,46]. They include patient demographic (age, race, marital stats), ADLs before cancer diagnoses, comorbidities (Diabetes, Hypertension, Arthritis, etc.), cancer characteristics (tumor size, staging, etc.), surgery and treatment indicators. The ICD-9 codes in the final dataset were mapped to the UMLS CUIs. These CUIs were used to extract the hierarchical relationships (Parents and Children) from UMLS until a common ancestor was found. The extracted hierarchies were added to the set of input variables.

AQ21 software was used to investigate the method with and without using background knowledge from UMLS. Application of the AQ21 software to SEER-MHOS data mapped to UMLS resulted in a number of models (rulesets) for predicting patients' deficiencies in performing activities of daily living. AQ21 has been executed in Approximate Theory Formation Mode (ATF), with weight w=0.3 of completeness vs. consistency gain. In the ATF mode, AQ21 produces rules that may be partially incomplete or inconsistent in order to maximize the rule quality measure. Below are two sample rules generated by AQ21 with and without using background knowledge:

Sample 1: AQ21
[Bathing impairment] <== [Race = 2,1,4: 70, 245, 22%]
[Hispanic = 2: 64, 241, 20%]
[Smoking = 2,3: 68, 238, 22%]
[Surgery = 51,40,27,0,45: 45, 113, 28%]
[Histology = 2,4,5,15,8,9,1: 74, 252, 22%]
[Stage = 0,1,2: 69, 244, 22%]
[Primary site and morphology = C0153458, C0153492, C0153532, C0242787, C0949022, C0235653, C0153483, C0153611, C0153555, C0153435, C0346782, C0153491, C0153612: 30, 34, 46%]
: p = 22, n = 2, q = 0.642

Sample 2: AQ21 with background knowledge
[Bathing impairment] <== [Race = 1,4: 64, 219, 22%]
[Hispanic = 2: 64, 241, 20%]
[Smoking = 2,3: 68, 238, 22%]
[Surgery = 32,51,40,0,45: 40, 95, 29%]
[Histology = 2,5,15,8,9,1: 68, 229, 22%]
[Cancer site = 2030, 25010, 21047, 21052, 21100, 29010,26000, 22020: 61, 169, 26%]
[Primary site and morphology = C0154077, C0007102, C0153532, C0005684, C0153555, C0024624, C0006142, C0235652, C0864875, C0346647, C0345921, C0242379, C0346629, C0345865, C0242788, C0034885, C0007107, C0345713, C0587060, C1263771: 38, 49, 43%]
: p = 23, n = 2, q = 0.653

The rules are similar using AQ21 with and without background knowledge. However, the quality of rule, as measured by Q(w), generated by the second method is improved. The last line in the rule set describes the numbers of positive examples (p), negative example (n) covered by the rule, and the rule quality. While the rules presented above correspond to each other, AQ21 with and without ontology are not guaranteed to create similar rules. Instead the program applies beam search to go through space of possible combinations of attributes and values to find the highest quality rules. Presence of additional ways of generalizing data available in the presence of hierarchies derived from an ontology may steer the process in different direction. Consequently, the quality of rules improves because of the ability to generalize using hierarchies.

2.3 Using non-IS-A Relationships

The current version of UMLS includes 727 types of relationships (NLM, 2016), with IS-A being just one of them. Semantics of these relationships need to be encoded in learning software, and their use and effect on reasoning process depends on specific meaning of that relationship. One simple way of encoding these relationships is that the data can then be extended by additional dimensions that correspond to presence of meaningful relationships. The following process is used to search for additional attributes to be added to problem representation space. It checks all pairs of attributes in the data and their values for existence of relationships. The following steps describe the method in terms of using UMLS, but can be easily extended to other ontologies.

1. Map the used attributes in the dataset (e.g., ICD-9) to the corresponding UMLS concept unique identifiers (CUIs). This can be done automatically if the coding system used to the attributes in data is part of UMLS' source vocabularies. Otherwise it can be done manually by experts.
2. For each pair of concepts retrieved:
   2.1. Search UMLS for non-hierarchical relationship(s) between the two concepts and all their parents (generalize using IS-A to the closest parents, or those within a predefined distance)
   2.2. Add all found relationships to a list of candidate attributes, and add new attributes to the data that indicates presence of the relationship. An example is shown in Figure 3. Concept X and Y exist in the dataset (see Figure 3a). The non-hierarchical relationship between X and Y is extracted from UMLS (see Figure 3b) and added as a new attribute X_Y to the dataset (see last column in Figure 3a). If the X and Y present in a patient's record, the value of the new attribute X_Y should be 1. Otherwise, it should set to 0.
   2.3. Apply attribute selection methods to filter out potentially large number of new relation-based attributes from the data. Those methods select attributes for learning by computing the discriminatory power of each attribute and comparing it with the acceptance threshold. Attributes whose discriminatory power is below the threshold will not be used for learning.
3. Apply standard learning algorithms on the data. At this point the data consists of original attributes including those mapped to UMLS and additional binary attributes that represent relationships. This encoding allows the use of any learning methods applicable to the original dataset.

The above method generates potentially very large number of new attributes, significantly increasing size of representation space for learning. The data are also typically sparse because of frequency of the related concepts co-occur in the data. Thus, efficient attribute selection methods (Step 2.3) need to be applied to reduce dimensionality. Experimental results performed on MIMIC III data indicate that even for large datasets there is a need to select most relevant attributes.

(a) An example

Patient ID	Code X	Code Y	X_Y
1	1	1	**1**
2	0	1	**0**

(b) Relationship in UMLS

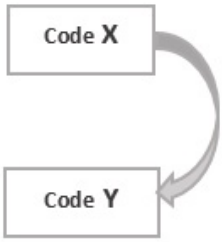


Fig. 3. Example candidate attribute

MIMC III contains rich clinical data including diagnoses and treatments for critical care patients. For example, some patients diagnosed with respiratory tract disease were treated by prednisone. As depicted in Figure 4, the non-IS-A relationship between "respiratory tract disease" and "prednisone" were extracted from the UMLS. Prednisone is a drug that "may treat" respiratory tract disease. Not surprisingly, the ontology includes also the reverse relationship "may be treated by". According to the Step 2.2, a new relation-based attribute "respiratory tract disease - prednisone" was added to the data. The value of the new attribute should be "1" for those patients for whom prednisone was used to treat that condition.

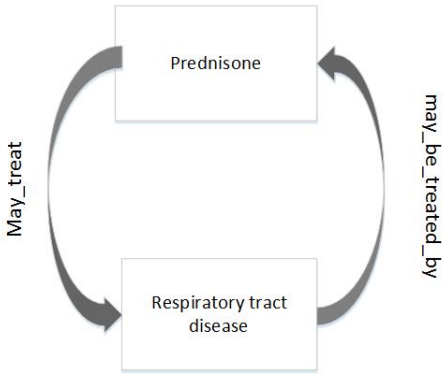


Fig. 4. Example non-IS-A relationship within MIMIC III data extracted from UMLS.

In order to test the described method for using non-IS-A relationships, it was applied to MIMIC III dataset in order to construct models for predicting 30-day post-hospitalization mortality. The performance of the two methods, with and without using semantic (non-IS-A relationships), were compared. The primary input attributes included the diagnoses registered during hospitalization. In order to prepare data, patients with age 65 and older and their admission records were extracted from the data (25,525 hospitalization records). Diagnoses originally coded with ICD-9 codes in the data were mapped

to Clinical Classifications Software (CCS) codes [48] to reduce the number of features in the data and group conditions into clinically meaningful categories.

As illustrated in Figure 3, candidate relationships were added as new attributes to the input set. CCS codes were mapped to CUIs. The IS-A relationships in UMLS is then followed to create the neighborhood space (parents and children) for each CUI. The neighborhood was used to find all non-IS-A relationships between each paired concepts. As a result, this method generated a relatively large number of new attributes (443) compared with the sample size. Hence, feature selection methods were used to decrease the size of representation space. The final dataset containing 421 features was used to learn models for predicting mortality.

Finally, standard learning algorithms such as Bayes network, naïve Bayes and logistic regression were applied to the dataset with and without using semantic. After adding semantics to the method, most models were able to capture more true positive cases and achieved higher recall which can be crucial in case of mortality prediction. For example, the naïve Bayes method without using semantics correctly classified 263 out of 555 death cases, while this method with semantics captures 10 more true positive instances.

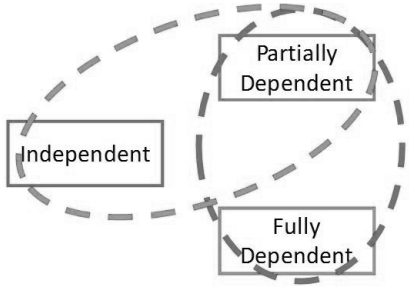
2.4 Learning Ordinal and Hierarchical Outputs

A typical approach to building multiclass classifiers is to learn models for each class against all other classes in the data as shown in Figure 5a. While effective in many cases, this approach suffers particularly when dealing with problems with many classes or when there is inherit structure to the concepts being described. Independent binary classifiers also do not allow for weighting types of mistakes made during classification (i.e., classification error of diabetes vs. cancer is worse than one of type I and type II diabetes).

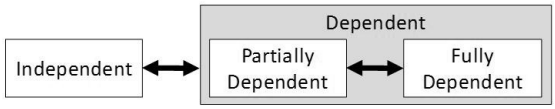
Building hierarchical classifiers has recently become popular approach in machine learning applications. Instead of building classifiers with large number of unrelated classes, information about structure of output (Figure 5c) may significantly improve performance of learning algorithms. Moreover, errors at lower levels of hierarchy are less critical than those at higher levels [42]. In order to build a hierarchical classifier, AQ21 starts with building models that distinguish general concepts at the top of the hierarchy (those connected to the root). Then data is limited to those within one general concept and models are built to describe sub-concepts. This operation is repeated recursively until all concepts in the hierarchy have corresponding models. AQ21 implements the method in breath-first search strategy but the order does not affect results nor computation time.

In addition to hierarchical structure within input attributes, AQ21 allows for ordered structures of output attributes. An example of ordinal (ordered) output is when one considers three or more levels of patient disability. A patient may be fully independent,

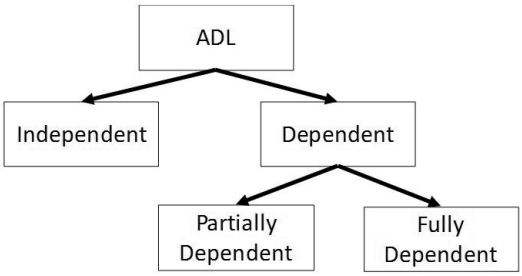
partially dependent/needs some help, or fully dependent in performing a certain task. In order to learn ordinal output, the system will first build a model to distinguish between fully independent patients and those with any level of dependence, and then among the dependent patients, distinguish between those with partial and full dependence, as illustrated in Figure 5b. It is clear that the order of values in the domain of ordinal attribute affects results of learning. It can be also easily observed that the order high-to-low will result is completely different classifier than low-to-high, thus one needs to carefully design attribute domains.



(a) Learning unordered output.



(b) Learning ordinal structure of output.



(c) Learning structured output.

Fig. 5. Learning unstructured and structured outputs.

3 Conclusions

Over the past decade significant progress has been made in the ability to use semantic information and ontologies in machine learning, despite being outside of mainstream research in the field. Research of our group at George Mason University focuses on selected aspects on using semantics (meta-values, ontologies, aggregated data) and is currently done in the context of analyzing medical, healthcare and health data.

In this study, we explored to include both hierarchical and non-hierarchical relationships in our data analysis. Both methods helped the learning process when models were created form the SEER-MHOS and MIMIC datasets. We have demonstrated that adding semantics to the ML method improved the performance of the prediction model by achieving higher recall. Capturing more true positive cases in the prediction model is important in some areas like predicting mortality or ICU admission. The results of the developed ML need to be interpreted by the domain experts. The new rules found by this study will be new hypotheses and validated by future investigations.

The presented preliminary study has a number of limitations. Several steps of data preparation need to be done manually, such as mapping data to the UMLS concepts. This causes problems related to resolving ambiguity between concepts and relationships among various ontologies, health agencies, users, and the way these ontologies are used. Different relationships between concepts may be important when a learning model preforms different tasks (i.e., differential diagnoses, comparative effectiveness of treatments, outcome prediction). Using UMLS in the ML is also challenge due to its extremely large size and high complexity. When we implemented the rich knowledge from UMLS and added them as new attributes, the dimension of the dataset increased dramatically. Thus, it brings another challenge for developing efficient attribute selection algorithms for data reduction. Currently our method increased only accuracy of models by small fraction. This may be the case that our simple strategy for hierarchical and non-hierarchical is not sophisticated enough to significantly improve the overall performance of our machine learning algorithm. Further research should be done to implement more sophisticated background knowledge and take greater advantage of the structure of the datasets. It is also important to study the effectiveness of problem of increasing dimensionality in context of size of data that is used for learning.

New wave of interest in artificial intelligence opens promise that methods using semantics and making computers “understand” objects which they reason or learn from will return to attention in machine learning. The most important work to be done in the near future concerns the ability to combine incredible advances make in statistical machine learning methods, with techniques described in this paper. High predictive accuracy of statistical models that also make sense to human experts is particularly important in domains such as healthcare where transparency is critical to users.

Acknowledgments
Our current activities on using semantic in machine learning are supported part by the Jeffers Foundation and LMI Academic Partnership Program.

References:

1. Gülçehre Ç, Bengio Y (2016) Knowledge matters: Importance of prior information for optimization. Journal of Machine Learning Research, 17(8), pp.1-32.
2. Tresp V, Bundschus M, Rettinger A, Huang Y (2008) Towards machine learning on the semantic web. In Uncertainty reasoning for the Semantic Web I (pp. 282-314). Springer Berlin Heidelberg.
3. Cai T, Giannopoulos AA, Yu S, et al. (2016) Natural Language Processing Technologies in Radiology Research and Clinical Applications. *Radiographics*, *36*(1), 176–191.
4. Wu, ST, Liu, H, Li, D., Tao, C., Musen, M. A., Chute, C. G., & Shah, N. H. (2012). Unified Medical Language System term occurrences in clinical notes: a large-scale corpus analysis. *Journal of the American Medical Informatics Association: JAMIA*, *19*(e1), e149–e156.
5. Xu R, Musen MA, Shah NH (2010) A Comprehensive Analysis of Five Million UMLS Metathesaurus Terms Using Eighteen Million MEDLINE Citations. *AMIA Annual Symposium Proceedings*, *2010*, 907–911.
6. Kassahun, Y., et al., Automatic classification of epilepsy types using ontology-based and genetic-based machine learning. Artificial Intelligence in Medicine, 2014. 61(2): p. 79-88
7. Bodenreider O (2008) Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearb Med Inform*: 67-79
8. Stearns MQ, Price C, Spackman KA, Wang AY (2001) SNOMED clinical terms: overview of the development process and project status. *Proceedings of the AMIA Symposium*, 662–666.
9. Hirsch JA, Nicola G, McGinty G, Liu RW, Barr RM, Chittle MD, Manchikanti L (2016) ICD-10: History and Context. AJNR Am J Neuroradiol, 37(4):596-9.
10. Huff SM, Rocha RA, McDonald CJ, et al. (1998) Development of the Logical Observation Identifier Names and Codes (LOINC) Vocabulary. *Journal of the American Medical Informatics Association: JAMIA*. 5(3):276-292.
11. The Gene Ontology Consortium. (2017) Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research*, *45*(Database issue), D331–D338.
12. Coletti MH, Bleich HL (2001) Medical Subject Headings Used to Search the Biomedical Literature. *Journal of the American Medical Informatics Association: JAMIA*, *8*(4), 317–323.
13. Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R (2011) Normalized names for clinical drugs: RxNorm at 6 years. *Journal of the American Medical Informatics Association: JAMIA*, *18*(4), 441–448.
14. Rosse C, Mejino JL Jr (2003) A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J Biomed Inform, 36(6):478–500.
15. Fragoso G, de Coronado S, Haber M, Hartel F, Wright L (2004) Overview and Utilization of the NCI Thesaurus. *Comparative and Functional Genomics*, *5*(8), 648–654.
16. Lindberg C (1990) The Unified Medical Language System (UMLS) of the National Library of Medicine. *J Am Med Rec Assoc*, **61**(5): p. 40-2.
17. Fung KW, McDonald C, Srinivasan S (2010) The UMLS-CORE project: a study of the problem list terminologies used in large healthcare institutions. *Journal of the American Medical Informatics Association: JAMIA*, *17*(6), 675–680.
18. Bodenreider O, Nguyen D, Chiang P, et al. (2013). The NLM Value Set Authority Center. *Studies in Health Technology and Informatics*, *192*, 1224.

19. Quinlan JR (1986) Induction of decision trees. Machine learning, 1(1), 81-106.
20. Fürnkranz J (1999) Separate-and-conquer rule learning. Artificial Intelligence Review, 13(1), 3-54.
21. Hearst MA, Dumais ST, Osman E, Platt J, Scholkopf B (1998) Support vector machines. IEEE Intelligent Systems and their Applications, 13(4), 18-28.
22. Lemeshow S, Sturdivant RX, Hosmer DW (2013) Applied Logistic Regression (Wiley Series in Probability and Statistics). Wiley.
23. Wang YX, Zhang YJ (2013) Nonnegative matrix factorization: A comprehensive review. IEEE Transactions on Knowledge and Data Engineering, 25(6), 1336-1353.
24. Schmidhuber J (2015) Deep learning in neural networks: An overview. Neural Networks, 61, 85-117.
25. Wang WY, Mazaitis K, Cohen, WW (2014) Structure learning via parameter learning. In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (pp. 1199-1208). ACM.
26. Kazakov D. and Kudenko D. (2001). Machine Learning and Inductive Logic Programming for Multi-agent Systems. In *Selected Tutorial Papers from the 9th ECCAI Advanced Course ACAI 2001 and Agent Link's 3rd European Agent Systems Summer School on Multi-Agent Systems and Applications* (EASSS '01), Michael Luck, Vladimír Marík, Olga Stepánková, and Robert Trappl (Eds.). Springer-Verlag, London, UK, UK, 246-272.
27. Kaelbling LP, Littman M, and Moore A (1996) "Reinforcement Learning: A Survey." *Journal of Artificial Intelligence Research*, 4:237-285, 1996.
28. Džeroski, S., De Raedt, L., and Driessens, K. (2001). "Relational Reinforcement Learning." *Machine Learning*, 43:7-52. The Netherlands: Kluwer Academic Publishers, 2001.
29. Tadepalli, P., Givan, R., & Driessens, K. (2004). Relational reinforcement learning: An overview. In *Proceedings of the ICML-2004 Workshop on Relational Reinforcement Learning* (pp. 1-9).
30. SEER-MHOS. http://healthcaredelivery.cancer.gov/seer-mhos/
31. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 101(23):e215-e220 [Circulation Electronic Pages; http://circ.ahajournals.org/content/101/23/e215.full]; 2000 (June 13)
32. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. *Scientific Data* (2016).
33. Michalski, R. S. and Larson, J., "AQVAL/1 (AQ7) User's Guide and Program Description," *Report No. 731*, Department of Computer Science, University of Illinois, Urbana, June 1975.
34. Wojtusiak J (2012) Recent Advances in AQ21 Rule Learning System for Healthcare Data, American Medical Informatics Annual Symposium, Chicago, November, 2012.
35. Wnek, J. and Michalski, R. S., "Hypothesis-driven Constructive Induction in AQ17-HCI: A Method and Experiments," *Machine Learning*, Vol. 14, No. 2, pp. 139-168, 1994.
36. Bloedorn, E. and Michalski, R. S., "Data-Driven Constructive Induction," *IEEE Intelligent Systems*, Special issue on Feature Transformation and Subset Selection, pp. 30-37, March/April, 1998.
37. Michalski, RS., "ATTRIBUTIONAL CALCULUS: A Logic and Representation Language for Natural Induction," *Reports of the Machine Learning and Inference Laboratory*, MLI 04-2, George Mason University, Fairfax, VA, April, 2004.

38. Wojtusiak J (2012) Semantic Data Types in Machine Learning from Healthcare Data, International Conference on Machine Learning and Applications (ICMLA), Florida, December, 2012.
39. Wojtusiak J, Michalski RS, Simanivanh T, Baranova AV (2009) Towards application of rule learning to the meta-analysis of clinical data: An example of the metabolic syndrome, International Journal of Medical Informatics, 78, 12, e104-e111.
40. Michalski RS, Wojtusiak J (2012). Reasoning with Missing, Not-applicable and Irrelevant Meta-values in Concept Learning and Pattern Discovery. Journal of Intelligent Information Systems, 39(1), 141-166.
41. Elixhauser A, Steiner C, Harris DR, Coffey RM (1998) Comorbidity measures for use with administrative data. Medical care, 36(1), 8-27.
42. Kaufman K, Michalski RS (1996) A Method for Reasoning with Structured and Continuous Attributes in the INLEN-2 Multistrategy Knowledge Discovery System, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, OR, pp. 232-237, August, 1996.
43. Amemiya, T., et al. (2007). Activities of daily living and quality of life of elderly patients after elective surgery for gastric and colorectal cancers. *Ann Surg*, 2007. **246**(2): p. 222-8.
44. Agborsangaya, C.B. et al. (2013). Health-related quality of life and healthcare utilization in multimorbidity: results of crosssectional survey: Qual Life Res, 22(4): p. 791-9.
45. Taneja, S.S. (2013) Re: impact of age and comorbidities on longterm survival of patients with high-risk prostate cancer treated with radical prostatectomy: a multi-institutional competing-risks analysis. *J Urol*, **189**(3): p. 901.
46. Vissers, P.A., et al. (2013) The impact of comorbidity on Health Related Quality of Life among cancer survivors: analyses of data from the PROFILES registry. *J Cancer Surviv*, **7**(4): p. 602-13.
47. Min H., Mobahi H., Vukomanovic S., Irvin K., Krasniqi I., Avramovic S., Wojtusiak J. (2016) "Applying an Ontology-guided Machine Learning Methodology to SEER-MHOS.
48. https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp